

THE UNIVERSITY of EDINBURGH

Text Technologies for Data Science

INFR11145

Web Search

Instructor:
Walid Magdy

4-Nov-2026

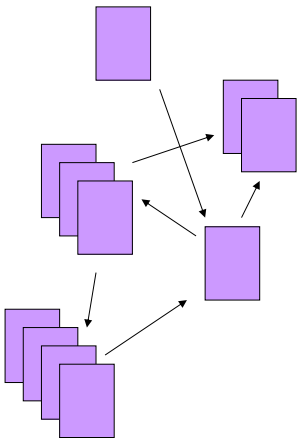
1

Lecture Objectives

- Learn about:
  - Working with Massive data
  - Link analysis (PageRank)
  - Anchor text

The Web Document Collection

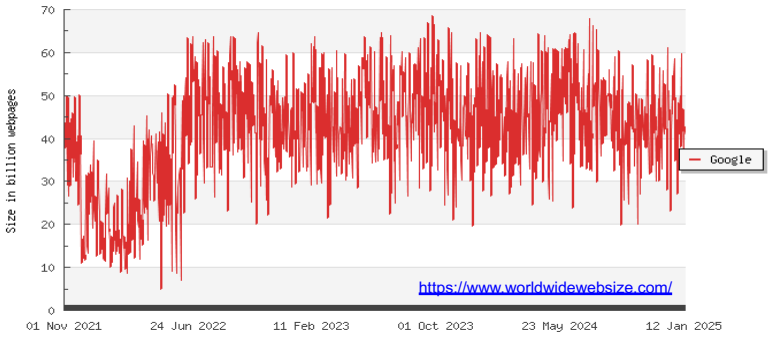
- Huge / Massive
- Graph / Connected
- No design/co-ordination
- Distributed content publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (DB) …
- Growing continuously!
- Content can be dynamically generated



The Web

Challenges

- Massive = 50-70 Billion webpages
- Growing
  - 20 PB/day in 2008 → 160 PB/day in 2013 → now??
  - 1 PB = 1,000 TB = 1,000,000 GB



Challenges

- No design/co-ordination
- Controlling quality!
  fake news, spam, generated content

- Challenging for a search engine
  - Technicalities: (collection, storage, processing, …)
  - Apparently relevant pages with low quality

- **Opportunity!**

What is the challenge in relevance?

- No clear semantics, contrast:
  - *"William Shakespeare"*
  - Author history's? list of plays? a play by him?
- Inherent ambiguity of language:
  - polysemy: "Apple", "Jaguar"
- Relevance highly subjective
- On the web: counter SEOs / spam

- **Potential solution: Wisdom of the crowds**

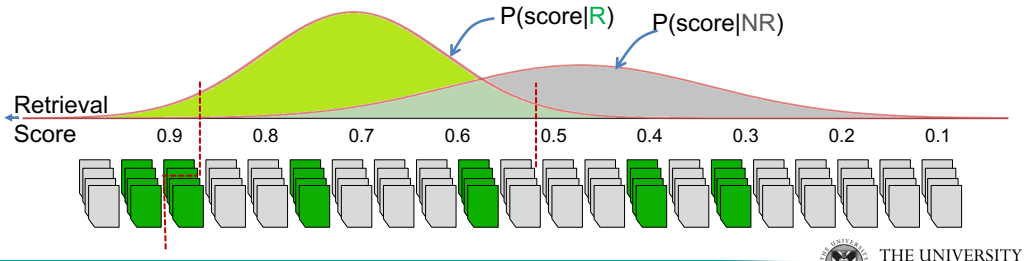
Effect of Massive data

- Challenges: storage, processing, networking, …
- Advantages: Makes stuff easier, how?

- Assume two good search engines the collects two sub-sets of the web
  - Search engine A collected N docs → precision@10 = 40%
  - Search engine B collected 4N docs → precision@10??

Effect of Massive data on Precision

- Assume two good search engines that collect two sub-sets of the web
  - Search engine A collected N docs → precision@10 = 40%
  - Search engine B collected 4N docs → precision@10??
    - Distribution of positive/negative scores stays the same
    - Precision/Recall at a given score stays the same
    - In any decent IR system: more relevant docs exist at the top → P@n ↑↑ ➔ precision@10 = 60% (increases)



Big Data or Clever Algorithm?

- For Web search, larger index usually would beat a better retrieval algorithm
  - Google Index vs Bing Index
- Similar to other applications
  - LLMs (now)
    - ChatGPT trained on whole internet (Common Crawl ~5 PB)
    - Larger data → more parameters → generally better performance
  - Google MT vs IBM MT (2006)
    - Statistical methods trained over 10x training data beat deep NLP methods with 1x training data
  - Question answering task (2002):
    - IBM Watson vs Microsoft experiment

Big Data or Clever Algorithm?

- Question answering task:
  - **Q**: Who created the character of Scrooge?
  - **A**: Scrooge, introduced by Charles Dickens in "A Chrismas Carol"
  - Requires heavy linguistic analysis, lots of research in TREC
- 2002, Microsoft
  - Identify (subj verb obj), rewrite as queries:
    - Q1: "created the character of Scrooge"
    - Q2: "the character of Scrooge was created by"
  - Search the web for exact phrase, get top 500 results
  - Extract phrase: ■Q1 or Q2■ , get most frequent ■
  - Very naive approach, ignores most answers patterns
  - Who cares!! Web is huge, you will find matches anyway

117	Dickens
78	Christmas Carol
75	Charles Dickens
72	Disney
54	Carl Banks
	…

Search “Microsoft”

Doc1

Microsoft.com

“Microsoft” mentioned 5 times

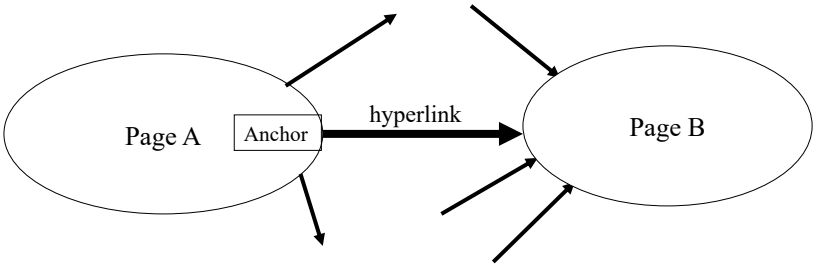
Doc2

Tutorial.com

Tutorial on MS word

“Microsoft” mentioned 35 times

The Web as a Directed Graph



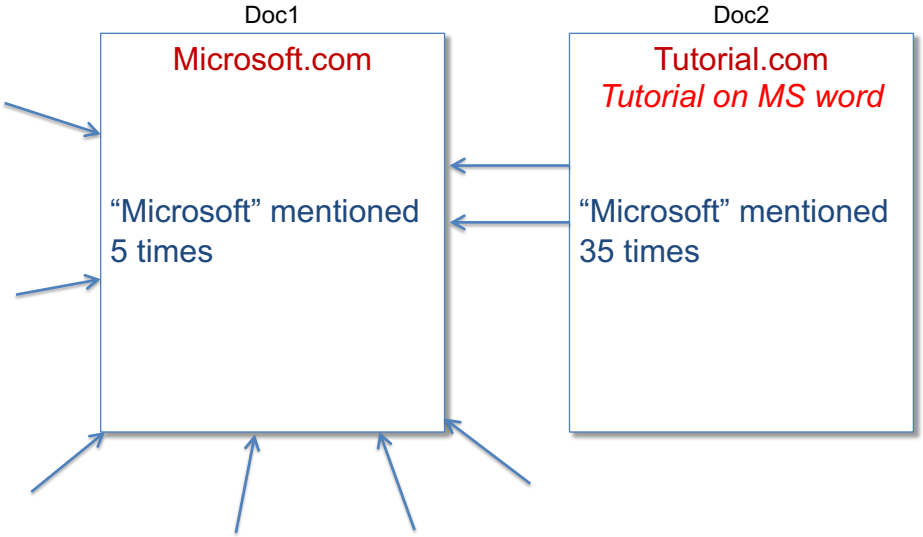
Assumption 1: A hyperlink between pages denotes author perceived relevance (quality signal)

Assumption 2: The text in the anchor of the hyperlink describes the target page (textual context)

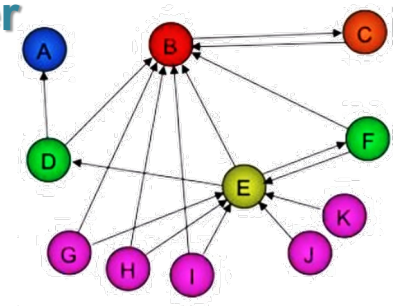
Links between Pages

- Google Description of **PageRank**:
  - Relies on the "uniquely democratic" nature of the web
  - Interprets a link from page A to page B as "a vote"
- A → B: means A thinks B worth something
  - "wisdom of the crowds": many links means B must be good
  - Content-independent measure of quality of B
- Use as ranking feature, combined with content
  - But not all pages that link to B are of equal importance!
    - Importance of a link from CNN >>> link from blog page
- Google PageRank, 1998
  - How many "good" pages link to B?

Search "Microsoft"



PageRank: Random Surfer



- Analogy:
  - User starts browsing at a random page
  - Pick a random outgoing link → goes there → repeat forever
  - Example: G → E → F → E → D → B → C
  - With probability 1-λ jump to a random page
    - Otherwise, can get stuck forever A, or B ↔ C
- **PageRank** of page x
  - Probability of being at page x at a random moment in time

PageRank: Algorithm

- Initialize $PR_0(x) = \frac{100\%}{N}$
  - *N*: total number of pages
  - $PR_0(A) = .. = PR_0(K) = \frac{100\%}{11} = 9.1\%$
- For every page x

$$PR_{t+1}(x) = \frac{1-\lambda}{N} + \lambda \sum_{y \to x} \frac{PR_t(y)}{L_{out}(y)}$$

- $y \longrightarrow x$ contributes part of its PR to x
- Spread PR equally among out-links
- Iterate till converge → PR scores should sum to 100%

- Initialize $PR_0(x) = \frac{100\%}{N}$
  - *N*: total number of pages
  - $PR_0(A) = .. = PR_0(K) = \frac{100\%}{11} = 9.1\%$
- For every page x

$$PR_{t+1}(x) = \frac{1-\lambda}{N} + \lambda \sum_{y \to x} \frac{PR_t(y)}{L_{out}(y)}$$

- $y \longrightarrow x$ contributes part of its PR to x
- Spread PR equally among out-links
- Iterate till converge → PR scores should sum to 100%

PageRank: Example

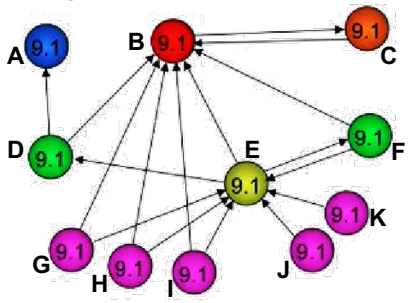
- Let λ = 0.82
- $PR_0(C) = PR_0(B) = \ldots = \frac{100\%}{11} = 9.1\%$
- $PR_1(C) = \frac{(1-0.82)}{11} + 0.82 \times PR_0(B)$

  $= 0.18 \times 9.1\% + 0.82 \times 9.1\%$

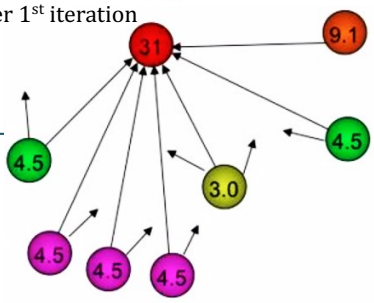
  $= 9.1\%$
- $PR_1(B) = \frac{0.18}{11} + 0.82 \times [PR_0(C) + \frac{1}{2}PR_0(D) + \frac{1}{3}PR_0(E) + \frac{1}{2}PR_0(F) + \frac{1}{2}PR_0(G) + \frac{1}{2}PR_0(H) + \frac{1}{2}PR_0(I)$

  $\approx 0.31 = 31\%$
- $PR_2(C) = \frac{0.18}{11} + 0.82 \times 31\% \approx 26\%$

Initial Pageranks

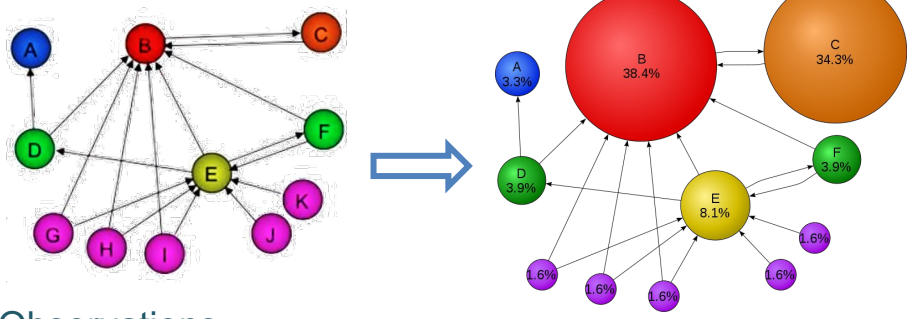


After 1st iteration



PageRank: Example result

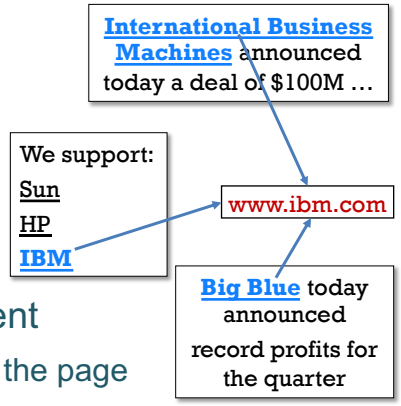
- Algorithm converges after few iterations



- Observations
  - Pages with no inlinks: PR = $(1-\lambda)/N$= 0.18/11 = 1.6%
  - Same (or symmetric) inlinks → same PR (e.g. **D** and **F**)
  - One inlink from high PR >> many from low PR (e.g. **C** vs **E**)

Anchor Text

- Anchor Text (text of a link):
  - Description of destination page
  - Short, descriptive like a query
  - Re-formulated in different ways
    - Human "query expansion"
- Used when indexing page content
  - Add text of all anchor text linking the page
  - Different weights for different anchor text
    - Weighted according to PR of linking page
- Significantly improves retrieval



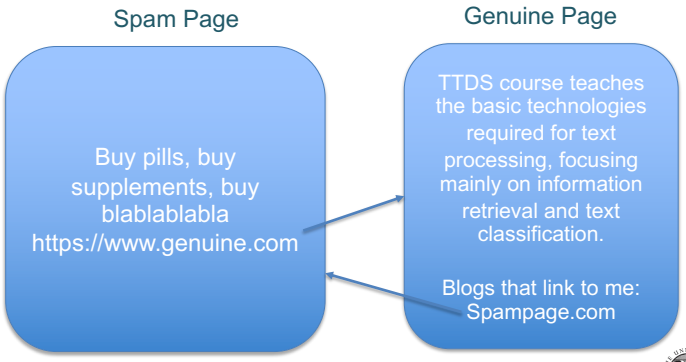
Vulnerabilities

- Hawthorne Effect: "Observation changes behavior"
- Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure"
- The Cobra Effect: "The solution worsens the problem due to other incentives"



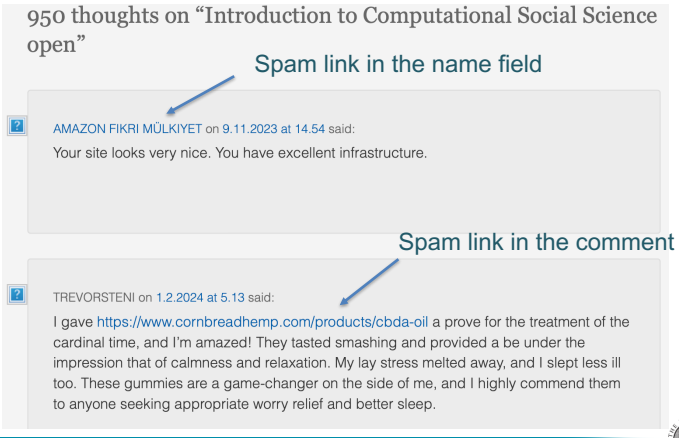
Trackback Links Spam

- "Blogs that link to me" feature
  - Pings the linked page (Sping)
  - Artificial feedback loops
    - Similar to "*follow back*" on Twitter



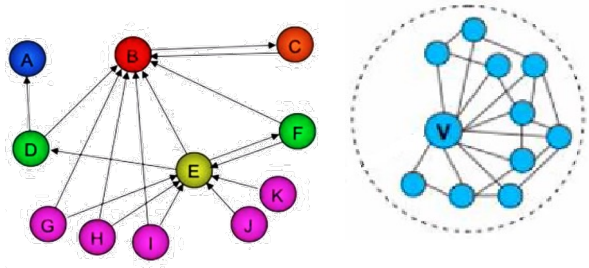
Comment Spam

- Links from comments on sites with high PR
  - One solution: insert `rel=nofollow` into links
    - Link ignored when computing PR



Link Farms

- Fake densely-connected graph (a clique)
- Hundreds of web domains / IPs can be hosted on one machine



- Applied on social media:
  Elmas T, Randl M, Attia Y. #TeamFollowBack: Detection & analysis of follow back accounts on social media. *ICWSM 2024*

Google Incident of 2003

Google Web Images Groups News Froogle Local more »
miserable failure Search Advanced Search Preferences

Web Results **1 - 10** of about **969,000** for **miserable failure**. (0.06 seconds)

Biography of President George W. Bush
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages
Past Presidents - Kids Only - Current News - President
More results from www.whitehouse.gov »

Monday, October 27, 2003
New Web Project
Let's get everyone to link to http://www.whitehouse.gov/president/gwbbio.html
with the words "Miserable Failure" Our goal is to make Shrubya the top google pick.

It's fun, it's easy just Miserable Failure
in your favorite web page will look like Miserable Failure

|Permalink| posted by George @ 10/27/2003 09:02:00 AM |

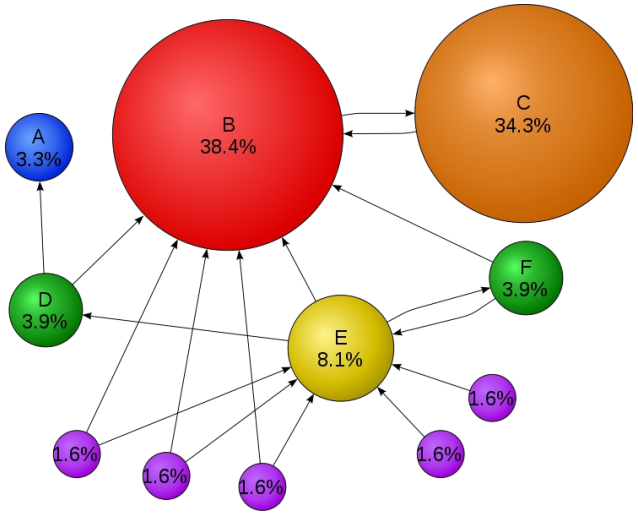
- "Anchor Text Spam"
- Fixed in 2007, when Google changed its index

Twitter Incident of 2020



25

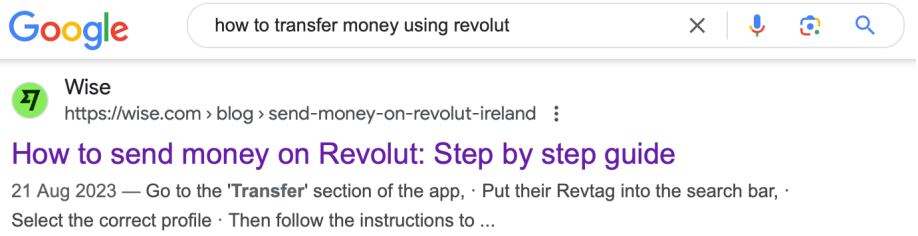
High Pagerank Websites Gaming The Algorithm



26

High Pagerank Websites Gaming The Algorithm

- Avoid giving outlinks
  - Sources can be cited without links
- Creating irrelevant content



- Stay tuned for more SEO tactics on the next lecture!

The Reality

- **PageRank** is used in Google, but is hardly the full story of ranking
  - A big hit when initially proposed, but just one feature now
  - Many sophisticated features are used
  - Machine-learned ranking heavily used
    - Learning to Rank (L2R)
    - Many features are used, including PR
  - Still counted as a very useful feature

Summary

- Web data is massive
  - Challenging for efficiency, but useful for effectiveness
- PageRank:
  - Probability than random surfer is currently on page x
  - The more powerful pages linking to x, the higher the PR
- Anchor text:
  - Short concise description of target page content
  - Very useful for retrieval
- Link Spam
  - Trackable links, link farms

Resources

- Text book 1: Intro to IR, Section 21.1
- Text Book 2: IR in Practice: 4.5, 10.3
- Page Rank Paper:
  Page, L., Brin, S., Motwani, R., & Winograd, T. (1999).
  The PageRank citation ranking: Bringing order to the web.
  Stanford InfoLab.
- Additional reading:
  - Dumais, S., Banko, M., Brill, E., Lin, J., & Ng, A. (2002)
    Web question answering: Is more always better?.
    SIGIR 2002.
  - Elmas T, Randl M, Attia Y.
    #TeamFollowBack: Detection & analysis of follow back accounts on social media.
    ICWSM 2024